タグの類似度に着目した利用者投稿サイト動画の多様性分析

上畑, 恭平
九州大学大学院システム情報科学府

伊東, 栄典
九州大学情報基盤研究開発センター

出版情報：IEICE Technical Report. 115 (381), pp.83-88, 2015-12-18. 電子情報通信学会
バージョン：

一般社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report

タグの類似度に着目した利用者投稿サイト動画の多様性分析

上畑　恭平†　　伊東　栄典‡

†九州大学システム情報科学府
‡九州大学情報基盤研究開発センター
812-8581 福岡市東区箱崎 6-10-1

E-mail:　†kamihata.k.411@s.kyushu-u.ac.jp,　‡ito.eisuke.523@m.kyushu-u.ac.jp

あらまし　近年，YouTube やニコニコ動画などの利用者投稿型動画共有サービスが人気である．これらのサイトは CGM (Consumer Generated Media) とも呼ばれる．現在 CGM サイトは社会に影響を与えるメディアに成長している．CGM サイトに毎日多数の動画が投稿されており，また膨大な利用者が動画を閲覧している． 現在，CGM サイトに投稿されるコンテンツの画一化が指摘されている．以前見たことのあるような動画や，派生動画が増えているように感じられる．コンテンツの多様性が減少して画一化が進むと，文化的多様性が失われると思われる．我々は，CGM サイトであるニコニコ動画を対象に，動画コンテンツの多様性動向を分析している．今回，動画に付与されたタグについて，動画タグ間の類似度に着目した分析を行った．その結果を報告する．

キーワード　CGM，コンテンツの多様性，タグ，cos 類似度，情報エントロピー

An analysis of movie contents diversity based on similarity of movie tags

Kyohei KAMIHATA†　　and　Eisuke ITO‡

†Department of ISEE, Kyushu University　6-10-1 Hakozaki, Higashi-ku, Fukuoka, 812-8581 Japan
‡Research Institute for IT, Kyushu University　6-10-1 Hakozaki, Higashi-ku, Fukuoka, 812-8581 Japan
E-mail:　†kamihata.k.411@s.kyushu-u.ac.jp,　‡ito.eisuke.523@m.kyushu-u.ac.jp

Abstract　Recent years, CGM (Consumer Generated Media) sites, such as YouTube nicovideo, become popular. CGM site also become a contents delivery media, which is able to give an influence on society. A lot of movies are posted to a CGM site every day, and also a large number of users are enjoying posted movies. At present, decreasing diversity of contents are indicated by some opinions. Posted movies may be similar with previous posted movies. The authors are afraid that decreasing diversity of contents causes less energetic cultural activity. In this paper, the authors tried to measure diversity of contents in a CGM site. They calculated the similarity between the movies using cosine similarity of tags of movie set.

Keywords　CGM，Contents Diversity，Tag，Cosine Similarity, Information Entropy

1. はじめに

近年，YouTube やニコニコ動画などの利用者投稿型動画共有サービスが人気である．これらのサイトは CGM (Consumer Generated Media) とも呼ばれる．サービス開始から数年経過した CGM サイトは，社会に大きな影響を与えるメディアに成長している．CGM サイトに毎日多数の動画が投稿されており，また膨大な利用者が動画を閲覧している．動画以外にも，小説投稿サイトや写真共有サイトも人気である．

我々は，ニコニコ動画を対象に，視聴者投稿コメントの感情分析に基づく動画ランキング手法の研究してきた[1]．また他の CGM サイトとして，小説投稿サイト「小説家になろう (syosetu.com) 」を対象に，小説に付与されたタグの分析や[2]，お気に入り登録の構造解析に基づく小説ランキング手法を研究してきた[3]．

現在，CGM サイトに投稿されるコンテンツの画一化が指摘されている．ニコニコ動画を運営するドワンゴ社川上量生氏へのインタビュー記事[4]では，再生回数上位の動画は，同一カテゴリの動画になりつつあるという傾向を指摘している．

コンテンツの多様性が減少し，画一化が進むと，文化的多様性が失われると思われる．ある特定の環境に特化し過ぎて多様性を失った文化からは，新たな文化的イノベーションが発生しにくいと思われる．

我々は，CGM サイトであるニコニコ動画を対象に，動画に付与されたタグの多様性を分析する事にした．本論文の構成を述べる．２節では国立情報学研究所が

提供するニコニコデータセットについて述べる．3 節では，動画集合における，様々な頻度解析について述べる．４節では，タグ多様性と情報エントロピーについて述べる．5 節では本論文の主題である，cos 類似度について述べる．6 節では，実データを用いたタグの cos 類似度の測定及び時系列での動向を示し，考察を行う．最後に 7 節で，まとめと今後の課題を述べる．

2. ニコニコデータセット

2.1. ニコニコ動画

ニコニコ動画は 2006 年 12 月 12 日にサービスを開始した，視聴者投稿型の動画配信サービスである．運営開始から 8 年経過した 2014 年 12 月末現在，1100 万件を超える動画が投稿されている． 会員数も膨大で，wikipedia[5]によると 2013 年 6 月時点での一般会員のアカウント数は 3000 万を超えており，有料のプレミアム会員数も 200 万を超えている．

2.2. ニコニコデータセット

国立情報学研究所は，情報学研究リポジトリと名付けた，研究用のデータ集合を提供している．ドワンゴ社および未来検索ブラジル社は，国立情報学研究所に協力して研究者にニコニコデータセットを提供している[6]．このデータセットにはニコニコ動画コメント等データと，ニコニコ大百科データが有る．本研究では前者の動画コメント等データを利用している．前者のデータ数などの概要を表 1 に示す．

ニコニコ動画コメント等データに含まれている項目の一部を表 2 に示す．

3. 動画の頻度分析

ニコニコデータセットの，動画メタデータを用いて，月ごとの動画投稿数，タグ数，頻度などを調査した．

3.1. 各月の動画投稿数

各月の動画投稿数を図 1 に示す．図 1 から，2007 年 3 月から 2012 年 11 月までの間，概ね右肩上がりに投稿動画数が増えていることが分かる．2012 年の動画の投稿数は月 18 万個程度である．

3.2. 一意なタグ数

次にその月に投稿された動画集合を対象に，付与されたタグについて調査した．図 2 に各月の一意なタグ数を示す．2008 年 3 月まで急激に増加し，その後は毎月 180 万個程度のタグ数になっている．

表 1　動画コメント等データ概要

項目	内容
期間	2007 年 3 月～2012 年 11 月
形式	JSON 形式
データ件数（動画数）	8,305,696
一意なタグ数	5,328,341

表 2　動画メタデータに含まれる項目

項目	説明
video_id	動画 ID
title	動画の題名
description	動画の説明文
upload_time	動画投稿日時
length	動画長
movie_type	動画のファイル形式
view_counter	閲覧回数（再生回数）
comment_counter	コメント数
mylist_counter	マイリスト登録数
tags	動画に付与されたタグ

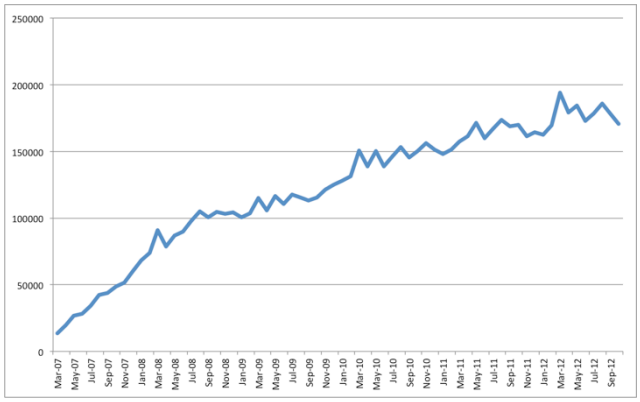

図　1 動画投稿数

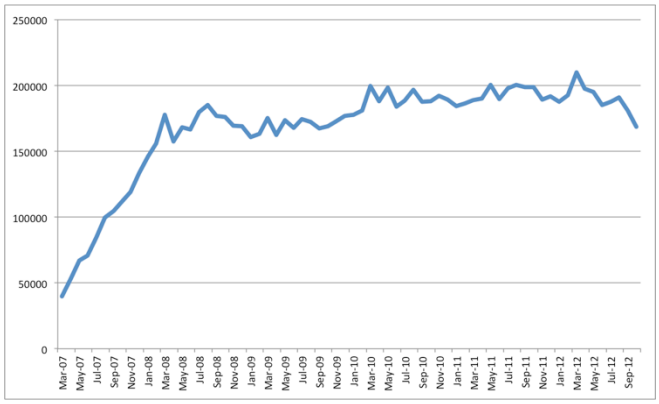

図　2　各月の一意なタグ数

3.3. 動画再生回数の順位-頻度

動画の再生回数を降順で並べたデータを作成した．そのデータに基づき，縦軸に再生回数，横軸に順位を取った散布図を図 3 に示す．なお，両軸とも対数尺度にしている．

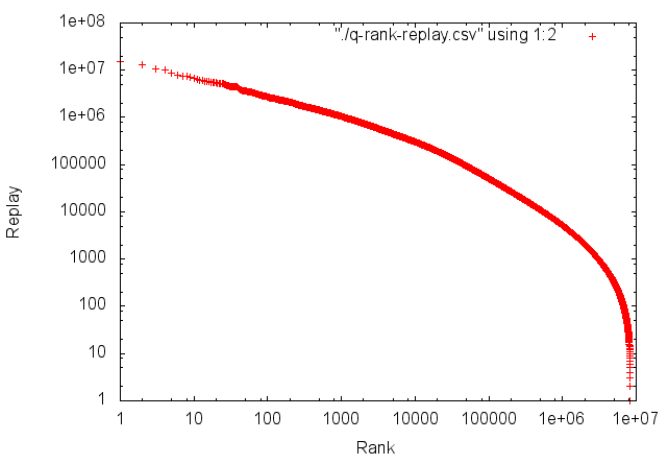

図 3　動画の順位-再生回数（対数尺度）

図 3 で分かるように，再生回数上位の動画の分布は直線に近い．両対数グラフで直線であるため，冪乗則（Power law）に近い分布をしている．しかしながら，再生回数の低い部分は，直線ではない．

次に，横軸に再生回数，縦軸にはその再生回数を持つ動画の数を散布図で描いた．この散布図を図 4 と図 5 に示す．

図 5 を見ると分かるように，横軸を対数尺度にすると，正規分布に近い曲線を描くことが分かる．このため，再生回数の分布は対数正規分布に近い分布であることが分かる．

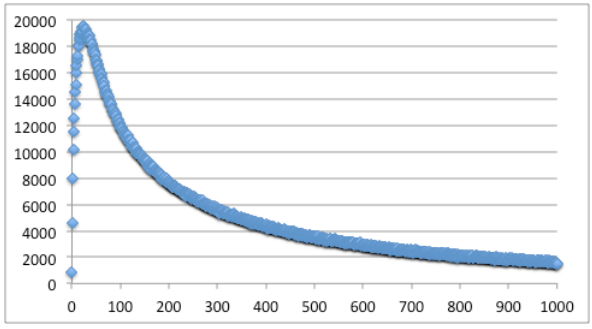

図 4　再生回数-動画数（再生回数 1000 回以下）

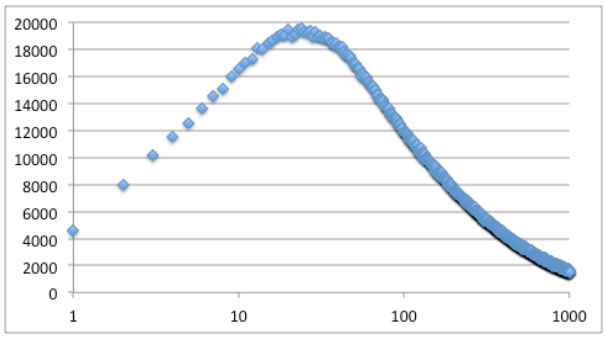

図 5　再生回数-動画数（再生回数 1000 回以下・横軸対数尺度）

3.4. タグ頻度（出現回数）の順位-頻度

動画に付与されたタグについて，各タグの出現回数（頻度）を降順で並べたデータを作成した．そのデータに基づき，縦軸に頻度，横軸に順位を取った散布図を図 6 に示す．なお，両軸とも対数尺度にしている．

図 6 の分布は両対数尺度で直線を示している．このためタグの出現頻度は冪分布であることが分かる．小説などの自然言語文における単語の出現頻度分布は冪分布になる．動画のタグ群も自然な分布をしていると言える．

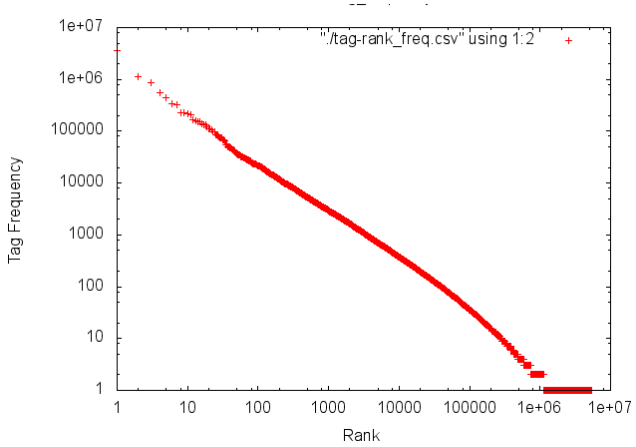

図 6　タグの順位-頻度（両対数尺度）

4. タグの多様性

先に，インタビュー[4]やブログ[5]で，CGM サイトへの投稿コンテンツの多様性減少への懸念が指摘されていることを述べた．筆者らの感覚としても多様性が減り，画一化が進んでいるように感じられる．本当に多様性が減少しているのかを判断するためには定量的な指標が必要である．

本研究では，コンテンツ多様性の度合いを数値で評価する指標を提案する．そのため，動画に付与されて

いるメタデータを，特に動画に付与された単語を用いて多様性の度合いを数値化する．

4.1. 多様性についての考え方

コンテンツの多様性について考えるため，最初に極端な場合を考える．もしもコンテンツが完全に画一化されているならば，全てのコンテンツに付与されるタグも同じになる．コンテンツ数（文書数）を n，タグの単語 w の文書頻度を $df(w)$とすると，全てのタグ w について $df(w)=n$ である．

逆に完全に多様であれば，全コンテンツに付与されるタグが異なるであろう．完全に多様な場合は，全てのタグ w について $df(w)=1$ である．

実際の動画では，カテゴリやジャンルを指定するタグを付与する．カテゴリタグは 30 個で有限であるため，これは多様にならない．また，図 6 で示したように，多くのタグは出現頻度が 1 である．頻度 5 以下のタグが殆どであるため，低頻度のタグだけを見て多様であるとすることは望ましくない．

4.2. タグ多様性の定義

情報エントロピーの考えを用いて，コンテンツ集合（文書集合）に対するタグの多様性を定義する．その際，以下の記号を用いる．

D : 動画集合，
n : 動画数 ($|D| = n$)，
W : タグ集合，
$df(w)$: タグ w の文書頻度．

情報エントロピーの考えた方を用いて，集合 D とタグ集合 W の多様度を単語当たりの情報エントロピー $H(W)$として定量化する．

$$H(W) = -\Sigma p(w)\log\big(p(w)\big),$$
$$p(w) = \frac{df(w)}{n}, \qquad 0 \leqq p(w) \leqq 1.$$

ここで $p(w)$ はタグ w の出現確率である．ニコニコ動画では，１つの動画に１つのタグを複数回付与できない．そのため，タグ w の出現確率は $p(w) = df(w)/n$ になる．

4.3. タグの多様性動向

情報エントロピーをタグに適用したタグ多様性の度合いでる $H(W)$ の値を，各月の投稿動画に付与されているタグで算出した．図 2 の青線は各月の $H(W)$の推移である．また，図 2 の赤線は，各月の動画集合で，動画に付与されたタグ集合における一意なタグの数を示す．

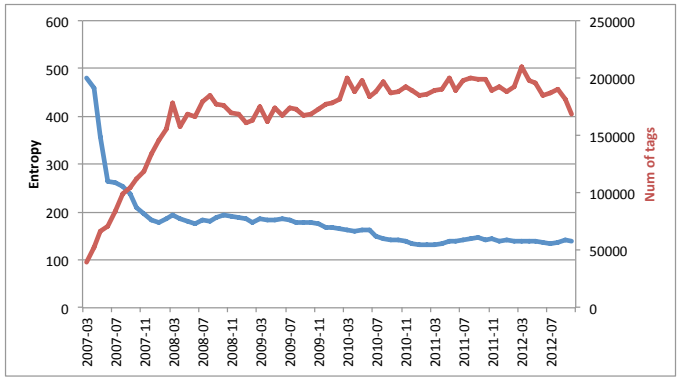

図 7　タグ多様度（青線）と一意なタグ数（赤線）

図 7 を見ると，一意なタグ数（赤線）は緩やかに増加しているのに対し，タグ多様度（エントロピー）は減少している．

5. 類似度と距離

投稿コンテンツの多様性減少について，各動画に付与されているタグ群同士の距離が近づいている，もしくは類似度が増加していることで判断できると考えた．

２つの集合の距離，類似度を測定する方法として，ユークリッド距離，マンハッタン距離，cos 類似度，ピアソンの相関関係，Jaccard 係数，Dice 係数，Simpson 係数などが知られる．

集合間の類似性を「共通要素が多く，非共通要素が少ない」場合に大きいとすると，先に述べた手法のうち cos 類似度，Jaccard 係数，Dice 係数，Simpson 係数を集合の類似性の指標として扱うことができる[7]．本研究では，この４つの手法のうち，比較的計算が容易であり，類似度の指標として最も用いられている cos 類似度を算出することにした．

5.1. cos 類似度について

cos 類似度とは，２つの文書間の類似度を測る手法の一つである．文書をベクトルとみなして，２つのベクトルの向きの近さを類似度の指標としたものが cos 類似度である．cos 類似度は０から１の値を取り，値が大きいほど２つの文書は似ていると言える．本研究では，各動画に付与されたタグ群をそれぞれ一つのベクトルとみなし，２つのタグ群の全ての組み合わせについて cos 類似度を算出し，それらを足し合わせることで cos 類似度の総和を求める．対象とする文書数を同じにして，cos 類似度の総和を比較することで，文書の類似度が増加しているかを判断できる．

5.2. cos 類似度の定義

以降で用いる記号を，以下のように定義する．

D	:	文書集合，
n	:	文書数 ($\|D\| = n$)，
W	:	タグ集合，
$d(\mathrm{w})_i$	:	動画 i に付与されたタグ群ベクトル．

cos 類似度を算出する際の２つのタグ群ベクトルを $d(w)_i$，$d(w)_j$とし．それぞれのベクトルが以下であるとする．

$$\mathrm{d(w)_i} = (\mathrm{a,b,c,d})$$
$$\mathrm{d(w)_j} = (\mathrm{a,c,e,f,g})$$

この時，$d(w)_i$と$d(w)_j$で次元が異なるので，次元を揃えたベクトル x を考える．

$$x = (a, b, c, d, e, f, g)$$

$d(w)_i$，$d(w)_j$から x の要素の有無を 0，1 で表したベクトル$d'(w)_i$，$d'(w)_j$を作成する．以下のようになる．

$$d'(w)_i = (1,1,1,1,0,0,0)$$
$$d'(w)_j = (1,0,1,0,1,1,1)$$

この２つのベクトル$d'(w)_i$，$d'(w)_j$を用いて cos 類似度を算出する．

$$Cosine\ Similarity = \frac{d'(w)_i \cdot d'(w)_j}{\sqrt{|d'(w)_i|} \cdot \sqrt{|d'(w)_j|}}$$

ニコニコ動画では，１つの動画に付与されるタグ数は最大１０個であり，かつ１つの動画に同じタグを複数回付与することはできない．そのため，

$$0 \leq |d(w)_i| \leq 10$$
$$0 \leq |d(w)_j| \leq 10$$

であり，$d'(w)_i, d'(w)_j$の要素はそれぞれ 0 か 1 となる．よって$d'(w)_i \cdot d'(w)_j$の値は，$d(w)_i$と$d(w)_j$に共通して含まれるタグの個数となる．

上記のように cos 類似度を，全ての文書集合 D の２つのタグ群の組み合わせについて算出し，それらを足し合わせた cos 類似度の総和を算出した．なお，組み合わせの数は $n(n\text{-}1)/2$ 個になる．

6. cos 類似度の総和の動向

動画に付与されたタグ群の cos 類似度の総和を月ごとに算出した．ここでは，文書数(動画数)は，$n = 1000$ である．図８では各月の再生回数上位 1000 個の動画を対象に cos 類似度の総和を算出した．図９では，各月の全動画のうちランダムに 1000 個選び，cos 類似度の総和を算出，これを 10 回繰り返し，10 回の平均値を算出した．

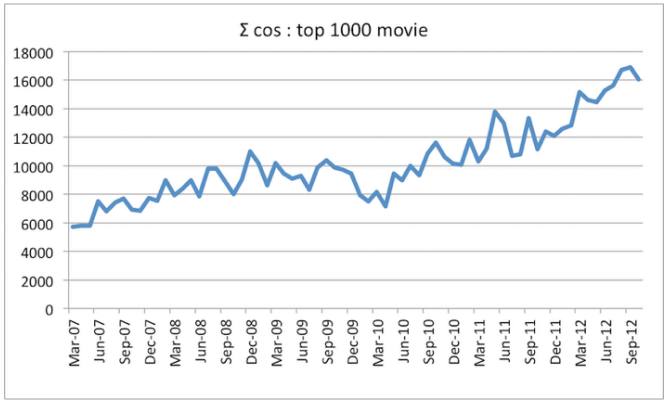

図８．cos 類似度の総和
(再生回数で上位 1000 個の動画)

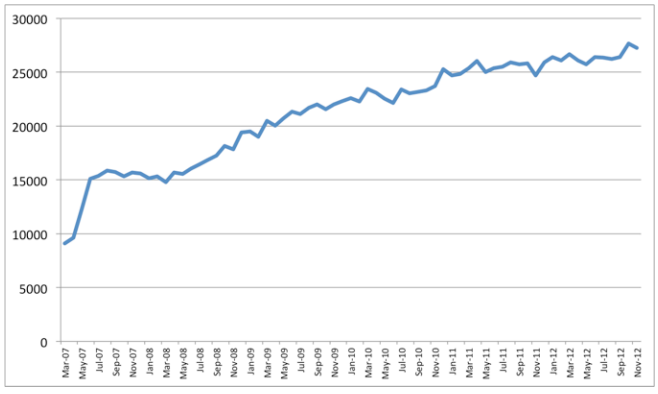

図９．1000 動画間の cos 類似度の総和
(ランダムに 1000 個の動画を選出)

図８，図９を見ると，どちらも cos 類似度の総和が緩やかに増加していることがわかる．

7. おわりに

本論文では，近年人気の CGM サイト，ニコニコ動画を対象に，コンテンツの多様性の動向を調査した．

情報エントロピーの定義を援用して，毎月のタグ多様性を数値で表現した．情報エントロピーを用いて，

月ごとのタグ多様性を算出し，それを時系列で折れ線グラフ表示した．その結果，一意なタグ数（赤線）は緩やかに増加しているのに対し，タグ多様度（エントロピー）は減少している．

また，投稿動画に付与されるタグについて，cos 類似度を用いて,毎月のタグの類似度を数値で表現した．cos 類似度を用いて,月ごとタグ群の cos 類似度の総和を算出し，それを時系列で折れ線グラフ表示した．その結果,再生回数上位 1000 個の cos 類似度の総和は緩やかに増加している．また，全動画のうちランダムに 1000 個選び，cos 類似度の総和を算出，これを 10 回繰り返し，10 回の平均値を算出したものについても，類似度の総和が緩やかに増加している．このことから，共通要素が多くなり，非共通要素が少なくなっている２つのタグ群の組み合わせが増加している．つまり，類似しているタグ群を持つ動画が増加している．

情報エントロピーと cos 類似度から，タグの多様性は徐々に失われている．この事は，投稿される動画の多様性減少を示すものと考えている．

今後は，タグ群についてクラスタリングを行い，タグ群の偏りや全体の傾向を調査していきたい．また，将来は電子コンテンツにおける利用閲覧モデルも考えたい．多様性喪失の原因として，端末の狭さがあると思われる．書店や図書館と異なり，PC 等では多数のコンテンツを一覧できない．また，コンテンツを試すには一つずつ閲覧するしかない．独力で多数を試すには時間が掛かるため，既知コンテンツに近いものを閲覧するのであろう．作者も，人気を得やすい分野のコンテンツを作りたがる傾向がある．利用者の閲覧モデルを作ることで，多様性喪失の原因が分かり，そこから多様性を保持する閲覧ソフトの開発ができると思われる．

謝　辞

本研究は JSPS 科研費 15K00451 の助成を受けたものである．

文　献

[1] Naomichi Murakami, Eisuke Ito: Emotional video ranking based on user comments, Proc. of iiWAS2011, pp.499-502, ACM, 2011.

[2] Eisuke Ito, Kazunori Shimizu: Frequency and link analysis of online novels toward social contents ranking, Proc. of SCA2012, pp.531-536, Nov. 2012.

[3] Kazunori Shimizu, Eisuke Ito, Sachio Hirokawa: Predicting Future Ranking of Online Novels based on Collective Intelligence, Proc. of ICDIPC2013, SDIWC, pp.261-272, 2013.

[4] Cakes, 川上量生:川上量生の胸のうち, https://cakes.mu/posts/5036 (accessed at Dec.12, 2014).

[5] ニコニコ動画 (Dec.12,2014) in *Wikipedia: The Free Encyclopedia*. Retrieved from http://ja.wikipedia.org/wiki/%E3%83%8B%E3%82%B3%E3%83%8B%E3%82%B3%E5%8B%95%E7%94%BB

[6] 国立情報学研究所，ドワンゴ社:ニコニコデータセット: http://www.nii.ac.jp/cscenter/idr/nico/nico.html, (accessed at Dec.12, 2014).

[7] Similarity and distance: http://wikiwiki.jp/cattail/?%CE%E0%BB%F7%C5%D9%A4%C8%B5%F7%CE%A5